

The University of Chicago

THE RELATION BETWEEN THE
DIFFICULTY AND THE DIFFERENTIAL VALIDITY
OF A TEST

A PART OF A DISSERTATION SUBMITTED TO THE FACULTY OF THE
DIVISION OF THE BIOLOGICAL SCIENCES IN CANDIDACY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

1936

By

MARION WEBSTER RICHARDSON

Private Edition, Distributed by
THE UNIVERSITY OF CHICAGO LIBRARIES
CHICAGO, ILLINOIS

Reprinted from
PSYCHOMETRIKA
Vol. 1, No. 2, 1936

The University of Chicago

THE RELATION BETWEEN THE DIFFICULTY AND THE DIFFERENTIAL VALIDITY OF A TEST

A PART OF A DISSERTATION SUBMITTED TO THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES IN CANDIDACY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

1936

By

MARION WEBSTER RICHARDSON

Private Edition, Distributed by
THE UNIVERSITY OF CHICAGO LIBRARIES
CHICAGO, ILLINOIS

Reprinted from
PSYCHOMETRIKA
Vol. 1, No. 2, 1936



THE RELATION BETWEEN THE DIFFICULTY AND THE DIFFERENTIAL VALIDITY OF A TEST

M. W. RICHARDSON

University of Chicago, Chicago, Illinois

Using scores of 1200 students on a long test as a criterion, each of five subtests of different difficulty has maximum correlation with the criterion when the criterion is dichotomized at a value appropriate to the difficulty of the subtest. A 50-item test element is scored on an all-or-none basis with different standards for passing, and the percentage of passes for successive points on the criterion variable is computed. The Constant Method is applied to this relationship. The limen thus computed is a measure of difficulty, the dispersion is a measure of average (or total) validity, and the slope of the curve is a measure of differential validity. The difficulty of a test element is thus directly related to the maximum differential validity.

I. EXPERIMENTAL DATA

The problem set in this investigation is the relationship between the difficulty of a test or test element and the degree of ability best discriminated by the test. The emphasis is properly placed upon the accurate description of difficulty and validity by the use of a simple mathematical function which will describe both in the same setting. Moreover, this function is to be brought into significant relationship with the statistical methods commonly employed in the interpretation of test results.

The test used in this investigation was the Co-operative General Culture Test, Form 1933.* The test is composed of 803 items, and requires 180 minutes. The choice of this test was based on the following requirements: (1) complete objectivity, (2) freedom from faulty items, (3) length, (4) possibility of getting a large sample from a representative larger population, (5) a wide distribution of difficulty of items. These requirements were approximately met by the Co-operative Test. The subjects were 1,200 college sophomores, a random sample from the 8,996 included in the norms provided by the Committee on Educational Testing. The data, on punched cards, were used at this time to eliminate faulty or completely invalid items. Using total score on the test as a criterion, the population was divided into twelve criterion groups of one hundred each. For each item, the percentage of correct response was computed for each of the twelve

*The 1933 College Sophomore Testing Program. *Report by the Committee on Educational Testing*. Washington: American Council on Education, 1933.

groups. These twelve percentages were plotted against the corresponding average criterion scores of the twelve groups. The slope of this empirical curve was taken as a crude measure of item validity. Items with flat or negatively sloping curves were not used in the experimental subtests. This impressionistic inspection of the validity of the items resulted in the elimination from the experimental tests of thirty or more items of doubtful characteristics. These items were allowed to remain in the criterion, their effect being assumed as negligible because of the length of the total test. From the 780 items remaining, five subtests were selected. The composition of these subtests is described in the following table:

TABLE I
COMPOSITION OF EXPERIMENTAL SUBTESTS

Subtest	Description	Percentage of Correct Answers	Number of Items
A	Very easy	78-95	50
B	Easy	60-77	50
C	Average	41-59	50
D	Difficult	23-40	50
E	Very difficult	5-22	50

The choice of items for the subtests was based on the following considerations, in order of importance: (1) absence from defect, (2) percentage of correct response, (3) balanced positions in the original test, so as not to give disproportionate emphasis to any type of material. To each of the 1,200 subjects a score was given upon each of the five subtests. This scoring was done by machine methods to insure a high order of accuracy.

The scores on the total test were taken as the "criterion" for purposes of this study. The raw scores were converted to normalized standard scores, by groups of 50 subjects. The criterion score of the fifty subjects in each of the twenty-four groups was considered to be the centroid as computed by the formula

$$\bar{x} = \frac{z_1 - z_2}{p_2 - p_1},$$

where $\bar{x}$ is the centroid of a truncated segment of the normal curve with unit standard deviation, z_1 and z_2 are the ordinates enclosing the segment at the left and right respectively, and p_1 and p_2 are the proportions of the area under the curve from the left to the ordinates

with the same subscripts. This procedure is merely a device for dividing the normalized criterion scores into a usable number of class-intervals.

II. PREDICTING TO A CRITERION OF TWO CATEGORIES

For many purposes we are interested chiefly in estimating from the test scores to which of two categories of some predicted variable the subject belongs. For instance, the testee is "qualified" or "not qualified" on an employment test; he is "passed" or "failed" on a test of achievement. The variance of criterion measures within each of the two categories is thus often neglected in practice. If the score meets a standard specified for the purpose of the test, it may not matter how much individuals, all meeting the standard, differ among themselves. The characteristic of the test which affords discrimination between individuals at different parts of the criterion scale will be referred to as "differential validity" in the subsequent discussion. It is pertinent to inquire the extent to which tests pitched at different levels of difficulty differ in the property of allocating the subject to two categories.

It might be expected that a sufficiently long test, homogeneous as to difficulty of test elements, would approach a condition of uniform differential validity for the total range of the criterion. This argument is based on probability theory, and there is not available any empirical evidence which could be used to test such a generalization. For a uniform difficulty of items which is not $p = .5$, the distribution of scores will be skew, for small and moderate values of n , the number of items. It would be possible, with much labor, to test empirically the degree of skewness of the distribution of scores of very long tests of homogeneous difficulty other than 50 per cent. It is not probable that the skewness diminishes to negligible values for tests of any practicable length. Associated with this skewness, there should be differences in the validity of test for predicting into two categories. It is possible to determine the degree to which the five tests of equal length but of varying difficulty are valid for predicting into two categories. The criterion was systematically divided into two categories at twenty-three different points indicated in Table II, and bi-serial r 's were computed for each. The choice of bi-serial r as a statistic for this purpose depends upon its special properties. The important property for our present purpose is that, for normal distributions of the continuous (non-dichotomized) variate, the value of bi-serial r is invariant with respect to the point of dichotomy. Another property

which must be noted in passing is that bi-serial r has limits of ± 1 only when the continuous variate is normally distributed.

Table II gives the values of bi-serial r for a large number of points of dichotomy of the criterion. These values are also represented in Figure 1.

TABLE II

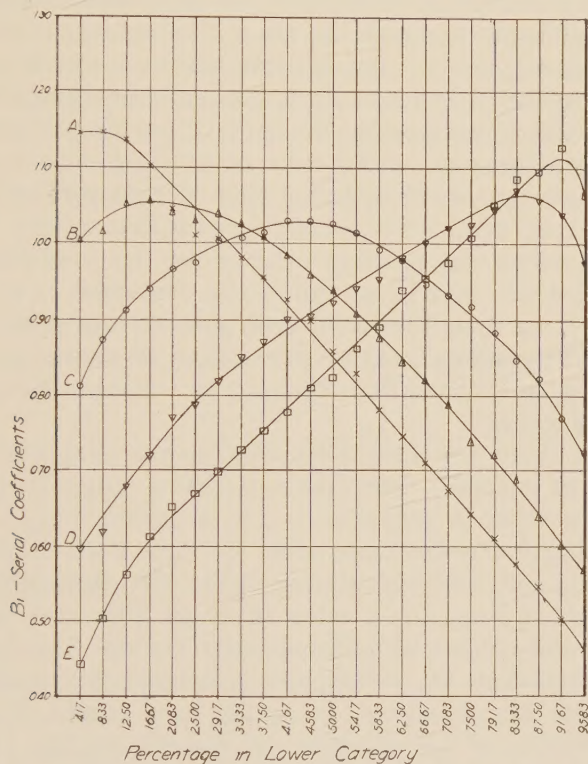
Bi-Serial Coefficients of Correlation of Subtests for Various Divisions of Criterion into Two Categories

Criterion		Coefficients				
Per-centage in Upper Category	Per-centage in Lower Category	A	B	C	D	E
4.17	95.83	.466	.570	.728	.976	1.066
8.33	91.67	.504	.600	.771	1.040	1.129
12.50	87.50	.549	.663	.823	1.059	1.096
16.67	83.33	.580	.689	.848	1.072	1.089
20.83	79.17	.611	.721	.884	1.046	1.051
25.00	75.00	.645	.744	.918	1.026	1.012
29.17	70.83	.675	.787	.934	1.022	.978
33.33	66.67	.711	.821	.949	1.001	.956
37.50	62.50	.746	.843	.977	.983	.940
41.67	58.33	.783	.875	.993	.953	.892
45.83	54.17	.826	.907	1.015	.941	.862
50.00	50.00	.860	.939	1.027	.924	.825
54.17	45.83	.900	.960	1.030	.904	.811
58.33	41.67	.927	.986	1.030	.900	.779
62.50	37.50	.957	1.012	1.015	.871	.752
66.67	33.33	.984	1.027	1.008	.850	.728
70.83	29.17	1.007	1.040	1.003	.818	.699
75.00	25.00	1.013	1.032	.978	.786	.670
79.17	20.83	1.048	1.042	.969	.768	.653
83.33	16.67	1.102	1.056	.940	.720	.613
87.50	12.50	1.139	1.054	.912	.679	.562
91.67	8.33	1.148	1.017	.871	.644	.503
95.83	4.17	1.148	1.007	.813	.597	.445

It is clearly seen from Figure 1 that Test A (the easy test) is most effective in separating off a small percentage from the lower part of the criterion distribution. Test E (difficult) is likewise most valid for separating off a small percentage of the criterion at the higher part of the distribution. The other tests have points or regions of maximum validity which are intermediate in the appropriate order. Test A is most valid for prediction to a two-categorized criterion which is divided at approximately -1.4σ . Test B is most valid for a criterion divided at approximately -1.0σ . Test C is most valid for a

criterion divided at -0.10σ . Test D is most valid for a criterion divided at approximately 1.0σ . Test E is most valid for a criterion divided at approximately $+1.4\sigma$.

These results prove, in a qualitative fashion, that fifty-item tests have pronounced differential validities appropriate to the various levels of difficulty. The objective of the next section is to outline in a



Bi-Serial Coefficients of Correlation for Various Dichotomies of the Criterion

FIGURE 1

more useful fashion a method of describing the difficulty of a test element and its validity in terms of a test discrimination function.

III. THE DISCRIMINATION FUNCTION

1. The Theoretical Curve

The difficulty of a test element which is objective may be simply described by the percentage of the population failing (or passing) the

item, i.e., by q or p . This definition of difficulty makes the one measure (p) primarily dependent upon the distribution of abilities of the arbitrary population. Obviously, an element or task which can be correctly performed by p_1 per cent of a given population may be correctly performed by a higher percentage p_2 of another population with a higher mean. This dependence of a measure of difficulty on the parameters of the distribution function of the population studied may not be escaped; difficulty is merely an obverse of achievement. When we say that a mental task is very difficult, for instance, we are merely expressing the fact that few individuals, of some group we have in mind, are capable of performing the task. If we use p (the percentage of passes) as a measure of difficulty of a test item, we understand that p is an average value; in fact p may be regarded as the average score of the group upon a test consisting of that one item. If we have a single task which has positive validity for the prediction of some criterion, and which may be unequivocally evaluated in terms of success or failure, we should have for each successively greater criterion value c_i a corresponding greater percentage of passes p_i . The relation between p_i and c_i could be regarded as a discrimination function for the single item.

However, there are various reasons for not using the single item for the present purpose. One reason is that a single item is subject to great fluctuations in response, i.e., it is unreliable. In the second place, the validity of most items is so low that item curves are comparatively flat, and unrepresentative of the test discrimination function. In the third place, the p value (average difficulty) of the single item is fixed and cannot be systematically varied. The discussion of the device adopted in the subsequent treatment will make clear the significance of this objection. In so far as a single objective item makes a prediction of the criterion, the distribution has a point character. The record on the item, taken at its face value, makes a total of pN times qN unit discriminations, where N is the population. In other words pN individuals are judged as better than qN other individuals. The item behaves as if pqN^2 all-or-none judgments were made upon the abilities of individuals of the group.*

Let us consider a test element to be sufficiently well represented by a limited number of items, homogeneous as to content and number

*Obviously, the number of possible judgments on N individuals with respect to all others is $\frac{1}{2}N(N-1)$. Not all judgments are made by one item, and the maximum number of judgments is $\frac{1}{4}N^2$ which occurs when $p=q=.5$. If we assume that item validity varies directly with the number of point discriminations made by the item, then the item of 50 per cent difficulty is the most valid, other things being equal.

Cf. Thurstone, Thelma Gwinn. "The Difficulty of a Test and Its Diagnostic Value," *Journal of Educational Psychology*, 1932, **23**, 335-43.

passing each. Furthermore, for convenience let the items each be passed by approximately 50 per cent of the population. These requirements are approximately satisfied by the fifty item test which has been designated as Test C. Although the test is to be considered as a unit, the constants of the distribution of scores on Test C and of its bivariate distribution against the criterion were computed for use in the subsequent analysis. The difficulty of this test element is now systematically varied by setting standards of difficulty at different points in the distribution of scores on this fifty-item test. This makes possible, for each argument of difficulty, the description of the discrimination function in terms of parameters which have meaning in themselves, and in terms of the more generalized psychophysical theory. In Figure 2 the bivariate distribution of the test element

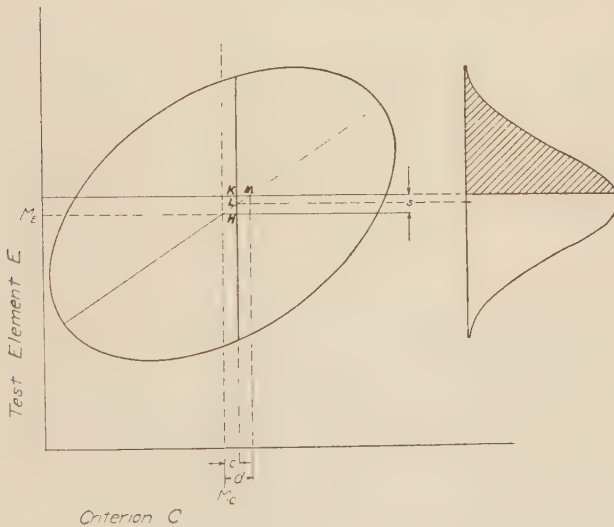


FIGURE 2

E and of the criterion C is represented in hypothetical terms. The means are M_e and M_c respectively. As a first approximation we assume linear regression, normal and homoscedastic E -arrays. Inspection of the obtained distribution suggests that this assumption is not unreasonable especially for values of E not too far from its mean. Let s , measured in standard deviation units from the mean of E , be the standard set for the difficulty of the test element, and constituting the operational determiner of difficulty of the element. One can express this difficulty of the element in terms of p , the percentage passing, but this is not immediately useful. This latter form of

expression becomes possible by reason of the fact that element E is dichotomized at s .

In Figure 2, the vertical line which contains the points K , L , and H represents any array in the bivariate distribution. The distribution of scores on the test element in this array is represented by the normal curve shown in Figure 2. In this array, the proportion of the group, all having a criterion score of c who exceed the standard s is indicated by the cross hatched portion of the curve at the right. This proportion varies with c , and will now be expressed as a function of c and the necessary parameters.

Let the deviation score in the element be designated by e .

Then the value of e corresponding to each c is given by the regression equation

$$e = b_{ec} c, \quad (1)$$

where b_{ec} is the coefficient of regression of e on c . For the array shown in Figure 2, e is the distance LH . Since L denotes the location of the mean of the distribution of the array, the array score which just equals the standard is represented by KL . This distance we shall call a . The expression for the value of the standard in the array is then

$$a = \frac{s - b_{ec} c}{\sigma_a}, \quad (2)$$

where σ_a is the standard deviation of the array, and is assumed to be constant for all arrays.

There will exist one array in which $a = 0$. This will occur when $s = b_{ec} c$. In this array, which can be represented by a vertical line passing through the point of intersection of the line of regression with the standard, the median array score just meets the standard. This point is M in Figure 2. We will now let the criterion score corresponding to this array be represented by

$$d = \frac{s}{b_{ec}}. \quad (3)$$

Then d is the parameter to be used in describing the difficulty of test element E when the standard is s . *The difficulty of a test element is defined as the standard score of the criterion measure at which the element (as a whole) is equally often passed and failed.*

Equation (2) may now be written, using the definition (3), thus:

$$a = \frac{b_{ec}(d - c)}{\sigma_a}. \quad (4)$$

According to the hypothesis adopted there will exist for each array, a theoretical proportion meeting or exceeding the standard of difficulty. The observed proportions in the twenty-four arrays will not agree exactly with the theoretical proportions. We will assume, consistently with the theory thus far applied to the problem, that the integral of the normal probability curve may be used as the theoretical curve. Clearly for a normal bivariate distribution, the proportions in the successive arrays vary with the corresponding criterion values exactly according to this function.*

If in equation (4) we now substitute $\sigma_d = \frac{\sigma_a}{b_{ec}}$, we have

$$a = \frac{d - c}{\sigma_d} . \quad (5)$$

The theoretical proportions are represented by

$$p = \frac{1}{N} \int_{\frac{d-c}{\sigma_d}}^{\infty} y \, dc , \quad (6)$$

where

$$y = \frac{N}{\sigma_d \sqrt{2\pi}} e^{-\frac{(d-c)^2}{2\sigma_d^2}} .$$

Following the usual procedure of minimizing the sum of the squares of the errors in the x (or base line) values corresponding to the proportions, we have as the expression to be minimized

$$v = \sum \left(\frac{d - c}{\sigma_d} - x \right)^2 , \quad (7)$$

the summation to extend over all arguments of the criterion for which data are available. The term $\frac{d - c}{\sigma_d}$ corresponds to the theoretical, and the x to the observed proportions. The formal problem is now to find values of d and σ_d which will make this sum of squares a minimum.

Differentiating equation (7) with respect to $1/\sigma_d$ and d/σ_d in turn, we have as normal equations

*The assumption of normality and homoscedasticity of arrays is tantamount to the choice of the integral of the normal curve as the theoretical function. The curve fitting procedure will hereafter be described in outline only, since it is adequately treated in other applications, especially in the literature on the Constant Method in psychophysics.

See Kelley, Truman L. Statistical Method. New York: Macmillan Co., 1924. Pp. 326-30.

$$\frac{1}{\sigma_d} \sum w c - \frac{d}{\sigma_d} \sum w - \sum w x = 0, \quad (8)$$

$$\frac{1}{\sigma_d} \sum w c^2 - \frac{d}{\sigma_d} \sum w c - \sum w x c = 0. \quad (9)$$

The w 's in the equations are the usual weights used in the Constant Method.

Solving (8) and (9) simultaneously,

$$\sigma_d = \frac{\sum w \cdot \sum w c^2 - (\sum w c)^2}{\sum w \cdot \sum w x c - \sum w x \cdot \sum w c}, \quad (10)$$

$$d = \frac{\sum w c \cdot \sum w x c - \sum w x \cdot \sum w c^2}{\sum w \cdot \sum w x c - \sum w x \cdot \sum w c}. \quad (11)$$

Five degrees of difficulty were arbitrarily created by setting up various standards for passing the test element. These standards, and the computed values of σ_d and d , are presented in Table III.

TABLE III.

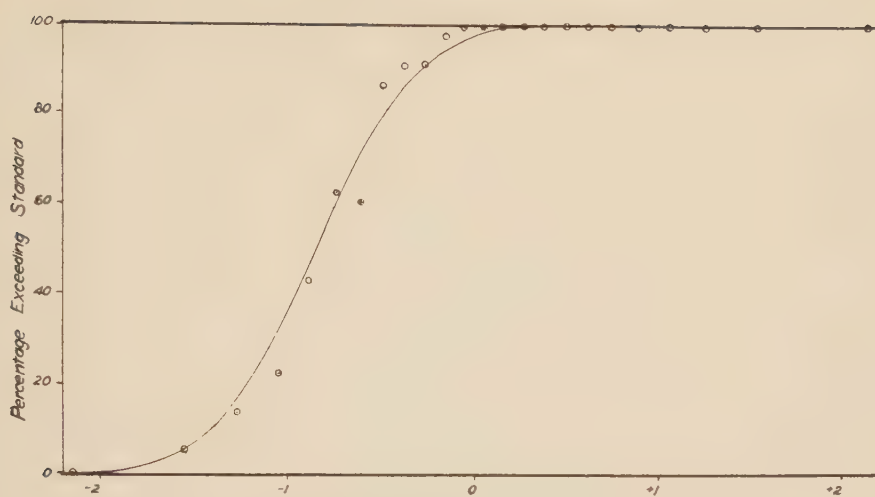
Parameters of Discrimination Function for Various Standards Subtest C

Standard in score Units of the Element	Standard in Standard Scores	Percentage of Passes p	Index of Validity σ_d	Liminal Difficulty d
37.07	1.0	22.0	0.550	0.92
31.59	0.5	34.6	0.412	0.34
26.11	0.0	51.8	0.316	-0.05
20.63	-0.5	64.9	0.317	-0.39
15.14	-1.0	78.2	0.433	-0.85

This table shows the fairly close relation between the difficulty of the test element and the standard adopted for passing the element.

In Figures 3 to 7 the observed proportions of each of the twenty-four criterion intervals who "pass" the element for the various standards of difficulty are shown, with the theoretical curve fitted to each.

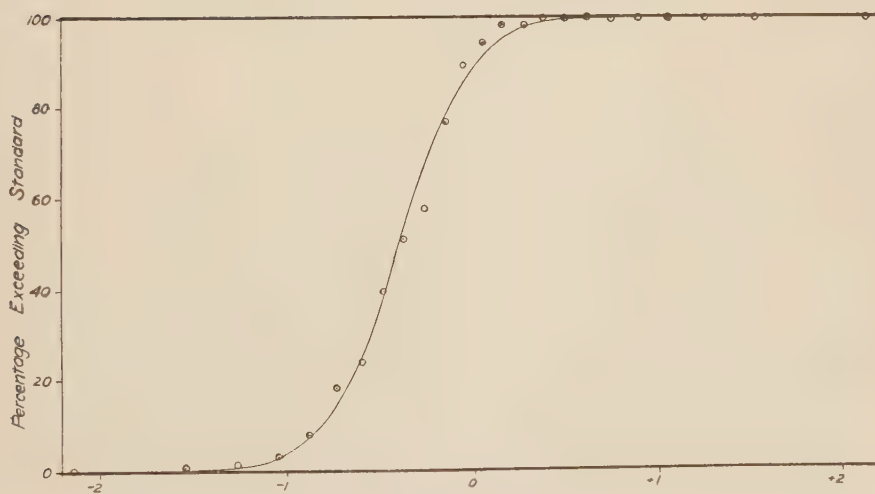
The test element, with standard set to allow 51.8 per cent of passes, has a difficulty measure of -0.05, measured in the standard units of the criterion. Under these conditions the test element is most valid for discriminating between individuals at or near the mean of the criterion group. The element does not differentiate between the abilities of individuals in the highest 20 per cent of the criterion group. The same is true of the lowest 20 per cent of the group. Considering that the region between the inflection points of the under-



Discrimination Function for Test Element C

FIGURE 3

Standard = $+1.0\sigma$



Discrimination Function for Test Element C

FIGURE 4

Standard: = $+0.5\sigma$

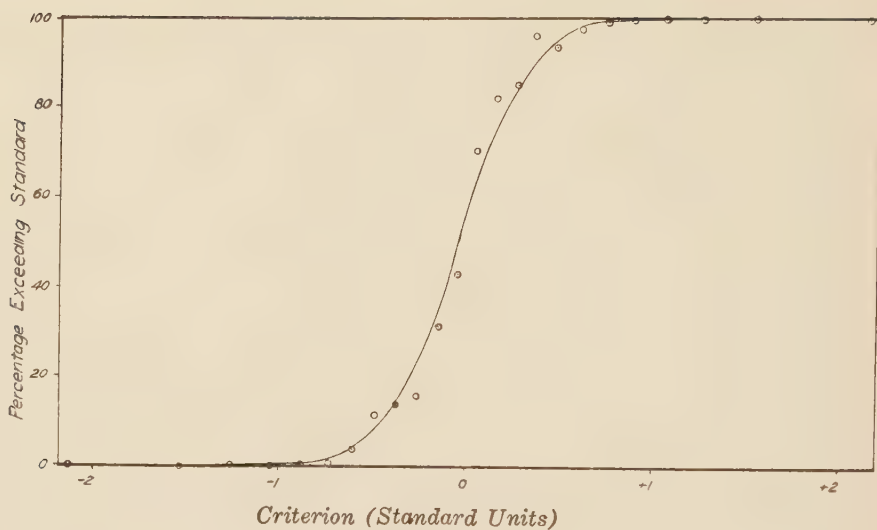


FIGURE 5
 Discrimination Function for Test Element C
 Standard: Mean = 0.0σ

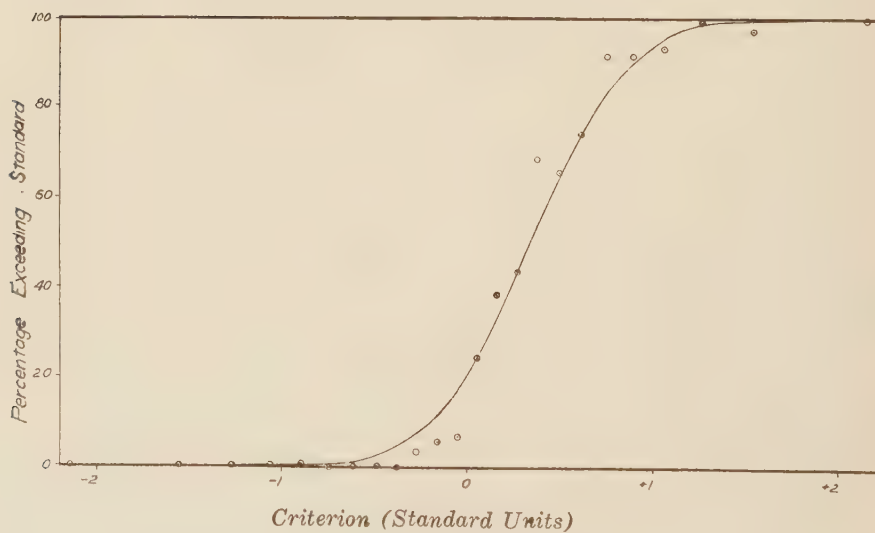
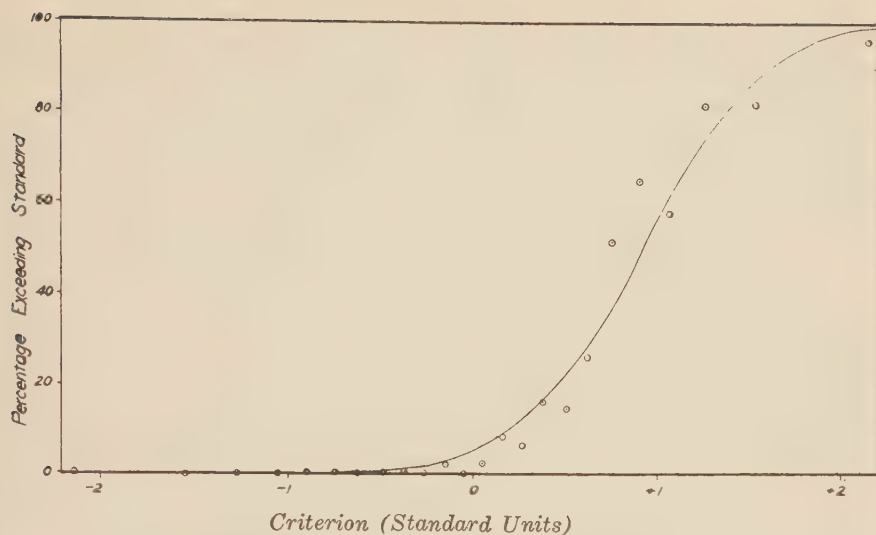


FIGURE 6
 Discrimination Function for Test Element C
 Standard: = -1.0σ



Discrimination Function for Test Element C

FIGURE 7

Standard: = -0.5σ

lying normal curve is to be regarded as the practical limit of the element's effectiveness, it can be said that a high validity element of this degree of difficulty is functional only between criterion scores of -0.37σ to 0.27σ . This is equivalent to saying that the element thus defined has validity for approximately the middle 25 per cent of the criterion group.

When the difficulty of the test element is raised so that the liminal difficulty is 0.34 (percentage of passes is 34.6), the area of effectiveness is in the direction of higher criterion scores. The region between one standard deviation above and below the liminal difficulty is defined by the limits 0.07σ and 0.75σ , which includes approximately 25 per cent of the criterion group. The measure of precision is the standard deviation of the theoretical curve. Its value is 0.412 in this case. When the standard of difficulty of the element is set at a still higher point, viz., 37.07 or 1σ in the element variable, the element is functional in discriminating between individuals only in the upper half of the criterion group. This element has a difficulty such that practically none of those in the lower half of the criterion group are able to perform the task set. The greater σ_d indicates that it is a less valid measuring instrument for the purpose of separating those with criterion scores above 1σ from those with scores below 1σ , than is the same element for standards of difficulty which will divide the crite-

tion group at points nearer to the mean. Likewise, standards of difficulty set for this element at points below the mean result in specific validity for a restricted sub-range of criterion ability, and practically no validity for other sub-ranges of ability. Likewise the element predicts the criterion less well in general as the standard departs from a liminal difficulty of 0σ .

IV. RELATION TO COEFFICIENT OF VALIDITY

In order to simplify equation (4), the substitution of σ_d for $\frac{\sigma_a}{b_{ec}}$ was made. This substitution was not arbitrary. We may write from ordinary correlation theory,

$$\sigma_a = \sqrt{1 - r_{ec}^2} \cdot \sigma_e, \quad (12)$$

and

$$b_{ec} = r_{ec} \frac{\sigma_e}{\sigma_c}. \quad (13)$$

Then we may rewrite (5) by substitution of (12) and (13), and simplify it to

$$\sigma_d = \frac{\sqrt{1 - r_{ec}^2}}{r_{ec}}, \quad (14)$$

when $\sigma_c = 1$. When (14) is solved for r_{ec} , we have

$$r_{ec} = \sqrt{\frac{1}{1 + \sigma_d^2}}. \quad (15)$$

Equation (15) gives us a method of estimating the validity coefficient. In Table IV is presented the validity coefficients, as estimated by equation (15), and as directly computed.

TABLE IV

Comparison of Measures of Validity of Elements of Varying Difficulty

Element	Validity Coefficient		Standard Error of (2)
	(1) As Estimated from the Discrimination Function	(2) As Computed Directly	
C; Standard: 1.0σ	0.876	0.896	0.0057
C; Standard: 0.5σ	0.924	0.931	0.0038
C; Standard: $0. \sigma$	0.954	0.958	0.0024
C; Standard: -0.5σ	0.953	0.954	0.0026
C; Standard: -1.0σ	0.918	0.941	0.0033
D; Standard: Median	0.892	0.902	0.0054

It is seen from the table that the validity of a test element as estimated from the dispersion parameter of the discrimination curve is a fairly close approximation to the validity as computed in correlational terms. This indicates that the assumptions back of the rationale were fairly justified. Moreover, in situations where such validity coefficients are not available, they may be easily estimated from the precision parameter of the psychometric curve which we have called the discrimination function.

V. DISCUSSION AND CONCLUSIONS

1. Although not directed at precisely the same problem, the present investigation confirms the conclusion of Thelma Gwinn Thurstone that a test composed of items of 50 per cent difficulty has a general validity which is higher than tests composed of items of any other degree of difficulty.*

Her results, however, are not inconsistent with the hypothesis that a test or test element of a difficulty other than the optimum might have a satisfactory differential validity which is confined to specific intervals of the criterion measure. Thus a test or test element, while indubitably having only fair general or average validity, might conceivably be highly valid for differentiating between individuals on some specified part of the criterion measure. The maximum general validity of elements of 50 per cent difficulty might be easily interpreted in terms of a differential validity peculiar to each degree of difficulty. If we assume that each criterion score c can be best distinguished from adjacent scores by an element of difficulty d , each value of c being associated with an element of optimal difficulty d_c , then it is clear that when $d = 0$ ($p = 50$ per cent), the condition of maximum general validity is attained, since $d = 0$ is the best representative of a distribution ($d_1, d_2, \dots, d_n$) of different optimal difficulties for the various values of c . Any other value of d would produce a greater average squared deviation from the respective optima for the various criterion scores.

2. It is definitely established by these experiments that tests of different difficulty will predict to a two-categorized criterion with different degrees of effectiveness. If it is desired to separate off a minor proportion from the lower end of the distribution of criterion scores, then an easy test has much greater validity than have more difficult tests. Moreover, the smaller the minor proportion to be separated from the criterion group at the lower end, the easier should the

**Op. cit.*

test be. The converse situation applies to minor proportions to be marked off from the upper end of the criterion group. It is to be noted, also, that if the general validity of a test can be regarded as a sort of average of the various differential validities, then its general or average validity decreases as the division point for the maximal prediction into two categories departs from 0σ , the mean of the distribution of criterion scores.

3. The terms "differential validity" and "liminal difficulty" have been repeatedly used in this investigation. The term "differential validity" refers to the capacity of the measuring instrument in discriminating between various levels of ability. When the differential validity of a measuring device varies from one region or interval of ability to another, it becomes pertinent to inquire at what specific point this discriminative capacity is a maximum. The criterion score at this maximum we have defined as difficulty, or liminal difficulty. The term "liminal" is aptly used because of its close analogy to the limen of sensitivity found from psycho-physical experiments. The method of describing the test element as a discrimination function has at least two advantages. One advantage is that the same type of mathematical function can be used as in the conventional psycho-physical theory; a skew function can be easily adopted if the use of a third parameter is desirable by reason of small elements of extreme positive or negative degrees of difficulty. A more important advantage is that the parameters of the curve have meaning built into them by the manner of derivation. In fact, the meaning of the parameters as used in the test setting have been expressed, to a fair approximation, in terms of the customary measures of validity and difficulty.

A further practical advantage of the method of measuring and expressing difficulty is that graphic methods are available. With the use of probability paper, the liminal difficulty of a test element and its validity may be quickly obtained.

4. The procedures outlined in the foregoing definitely point to the unsatisfactory nature of the common practice in the construction of tests of letting difficulty take care of itself. It is true that a test, no matter how the difficulty of its elements is distributed, will give a distribution of scores which may have practical usefulness. We cannot assume, however, unless the test is exceedingly long, that a chance distribution of difficulty will give scores which are linear with true measures of the ability. If, on the other hand, we have some specific purpose of prediction in mind, we may utilize the differential validities of test elements to our advantage. Suppose, for example,

that a test of clerical aptitude is meant to sort out the best 15 per cent of all applicants. This is on the assumption that the labor market is such that one hundred persons will apply for fifteen positions. It is then clear that the optimal difficulty of test elements should be in the neighborhood of $+1\sigma$ and that easier tasks would give us discriminations between individuals in whom we are not interested. The proper choice of difficulty gives us the maximum discrimination between the applicants we care to consider seriously. The converse consideration would apply to any situation in which a minor proportion from the lower end of a distribution is to be separated off for any purpose. Under any circumstances involving educational or psychological measurement, the distribution of difficulty of the elements or tasks can be arranged to fulfill more accurately the purposes of the measurement.

